課題2の補足

担当教授：植野真臣, TA: 木下 涼

e-mail: kinoshita@ai.is.uec.ac.jp

0.1　全体の流れについて

課題 2 の目標は spam のテストデータ TD.csv の各行 (各メール) がスパムか (1 をとるか) 否か (0 をとるか) を (2.3) 式で分類し，全行 (全メール) に対する正答率を測定することである．
そのためには，パラメータ θ_{x_0} と θ_{ix_0k} を spam の学習データ LD.csv から推定しなければならない．
ここでは，パラメータを最尤推定量の式 (2.5) により推定する．
式 (2.5) を計算するには N, N_{x_0}, N_{ix_0k} をそれぞれ求めなければならない．
N は学習データのサイズ，すなわち LD.csv の行数である．
N_{x_0} は目的変数 (一番右側の列) が x_0 という値をとる回数であり，実装では N_0 と N_1 をそれぞれ求める．
N_{ix_0k} は目的変数が x_0 という値をとり，かつ各単語 i が k という値をとるような回数である．実装では，各単語 i に関して $N_{i00}, N_{i01}, N_{i10}, N_{i11}$ を求める．

以上をまとめると，課題 2 の NB.java では，関数”setFrequencyTable”で N_{x_0}, N_{ix_0k} を求め，それを用いて関数”getParameters”でパラメータ θ_{x_0} と θ_{ix_0k} を求め，そのパラメータを用いて関数”classification”で TD.csv の正答率を測定する．

0.2　分類精度のめやす

正しく実装すると，分類精度は 0.9 以上になる．分類精度が 0.869 になる人は，うまく学習できておらず，全てのテストデータで目的変数を 0 と予測していると思われる．